April 2014 Technical Report: KUL/ESAT/PSI/1403

Fast vocabulary learning for disordered speech vocal interfaces

Jort F. Gemmeke, Siddharth Sehgal, Stuart Cunningham

PSI

Kasteelpark Arenberg 10 – box 2441
B-3001 Heverlee, Belgium

KU LEUVEN

Telephone: +32-(0)16-32.17.13 Fax: +32-(0)16-32.17.23 http://www.esat.kuleuven.be/psi

Fast vocabulary learning for disordered speech vocal interfaces

April 25, 2014

Abstract

Over the past decade, several speech-based electronic assistive technologies (EATs) have been developed that target users with dysarthric speech. These EATs include vocal command & control systems, but also voice-input voice-output communication aids (VIVOCAs). In these systems, the vocal interfaces are based on automatic speech recognition systems (ASR). In this work we evaluate an alternative approach, which works by mining utterance-based representations of speech for recurrent acoustic patterns, with the goal of achieving usably recognition accuracies with less speaker-specific training data. Comparisons with a conventional ASR system on dysarthric speech databases show that the proposed approach offers a substantial reduction in the amount of training data needed to achieve the same recognition accuracies.

1 Introduction

Spoken language communication is central to daily life, but as many as 1.3% of the population cannot use natural speech to communicate reliably [1]. Impaired speech can often be unintelligible to unfamiliar communication partners, and it also can make the use of conventional voice controlled command & control (C&C) systems problematic. Such systems, however, can significantly contribute to the independence of living and quality of life of users with restricted motor control [2].

Over the past decade, several speech-based electronic assistive technologies (EATs) have been developed that target users with dysarthric speech. These EATs include vocal C&C systems [3, 4], but also voice-input voice-output communication aids (VIVOCAs) [5]. The three challenges these systems face is that 1) The number of phones that can be produced is often severely restricted, making it difficult to distinguish between words, 2) dysarthric speech varies greatly between speakers and 3) speaking often requires great effort, thus restricting the amount of training or adaption material that can be collected.

Conventional EATs for dysarthric speech are based on automatic speech recognition (ASR), employing either speaker-independent acoustic models trained on a large corpus with adaptation to the target speaker [6, 7, 8, 9], or speaker-dependent modelled trained directly on speech material from the target user [3, 5]. Although adaptation approaches typically require less speech material from the target user than speaker-dependent modelling approaches, their performance largely depends on the exact speech characteristics.

Table 1: Dysarthric speech databases used for evaluation. Intelligibility is denoted with E for less than 20% intelligibility, D for 20-50 % intelligibility and C for 50-90 % intelligibility. Starred labels are the result of informal listening tests, while non-starred labels are measured using the word-level intelligibility assessment procedure described in [5].

	VIVOCA													STARDUST		
Speaker	1	2	3	4	5	6	7	8	9	10	11	12	13	1	2	3
Vocabulary size	35	14	19	57	35	64	100	28	11	6	20	16	13	19	10	13
Total Utterances	1225	742	514	2956	1674	2821	4543	933	220	145	269	283	272	628	417	708
Intelligibility (%)	E*	D	E	E	E	E	E	E	C	E*	E	E*	D*	E	E	E

In this work we evaluate an alternative approach to conventional Hidden Markov Model (HMM)-based ASR, which works by mining utterance-based representations of speech for recurrent acoustic patterns [4]. This speaker-dependent approach, developed in the ALADIN project, maps these acoustic patterns directly to (parts of) commands, which means it is language independent and does not require a pre-defined vocabulary, grammar or even knowledge of word order in the training data.

Recent evaluations have shown the ALADIN system yields relatively high recognition accuracies even after a single training sample of each word or command [10, 11]. In this work, we investigate to what extent this approach can be used to augment, or even replace, a conventional speaker-dependent ASR system for dysarthric speakers. The goal is to achieve usable recognition accuracies with less training data, in order to minimize the initial effort of the target user.

The contributions of the paper are twofold. First, we characterize the performance of both a speaker-dependent ASR approach and the ALADIN approach as a function of the amount of training data. Compared to earlier evaluations of the ALADIN system on dysarthric speech, the databases employed in this work constitute much larger amounts of (possible) training data, and speech from more severely impaired speakers. Second, we evaluated the performance on both isolated words and on C&C sentence data, which allows us to investigate to what extend the ALADIN approach — learning from sentence data without strong supervision such as word order and vocabulary — impacts the recognition accuracy.

In Section 2 we briefly describe the ALADIN system. In Section 3 we describe the dysarthric speech databases used for evaluation. In Section 4 we describe the experimental setup, and we present our results in Section 5. We conclude with a discussion and directions for future work in Section 6.

2 ALADIN

2.1 Knowledge representation

Each spoken command, for example “turn on the television”, is associated with a possible action. A manual execution of the action would for example be pressing the standby button on the television remote control. Actions are represented using a *semantic frame* [12], a data structure that represents the semantic concepts that are relevant to the execution of the action and which end-users are likely to refer to in their spoken commands. Each semantic frame represents a

possible action, and is composed of slots, which in turn contain slots or values. To continue the example, a semantic frame could contain two slots, <device> and <action>, allowing the values <television, radio> and <on,off>, respectively.

Internally, a semantic frame description is represented as a binary *label vector* indicating the presence or absence for all possible slot-values collected over all frames and slots. Using the example semantic frame, the command "turn on the television" would be represented as [1 0 1 0].

2.2 Non-negative matrix factorisation

The ALADIN approach works by determining recurrent acoustic patterns in spoken commands, and is based on a non-negative matrix factorisation (NMF) approach [11, 10, 13]. NMF is a technique which decomposes a non-negative matrix into the product of two non-negative low-rank matrices [14, 15, 16, 17]. The system works as follows. The spoken command is converted into an utterance-based vector representation, the *acoustic representation*. In a nutshell, this representation is constructed by making a histogram of the co-occurrences of Gaussian posteriors over time, with the Gaussian acoustic model obtained in advance. The acoustic model is estimated through unsupervised k-means clustering of the training data, followed by estimating a single full co-variance Gaussian on each cluster.

The collection of spoken training commands is concatenated into a matrix, which is then factorised by NMF into a matrix representing recurrent acoustic patterns (the *dictionary*), and a matrix of activations of these patterns over the training utterances. This factorisation is guided (regularized) by the label vectors to ensure that the obtained acoustic patterns correspond to slot-values within semantic frames.

2.3 Decoding

Decoding an observed utterance entails using NMF to find the combination of dictionary elements needed to represent the acoustic representation of the spoken command. Through the correspondence of these activations with the slot-values in semantic frames, we infer a semantic frame description of the observed utterance: for each slot whose cumulative slot-value activations exceeds a threshold, we assign the value with the largest activation.

3 Speech material

In this work, we employ two datasets, VIVOCA (1 and 2) and STARDUST. The methods employed to collect this data are described in [5] and [3] respectively. All speakers had mild to moderate dysarthria. The speech was recorded directly onto either a laptop computer or a PDA hand-held computer.

3.1 VIVOCA

The vocabulary size, number of utterances and intelligibility assessment are shown in Table 1. The data from the VIVOCA project contains words that were used by the speakers to compose messages on voice output communication aid.

The size of the vocabulary for each speaker varied according to the message building method the speaker choose to use (see [5], section II B). For each speaker the message building method, and the input and output vocabularies were individually tailored to the needs and wishes of each participant. Generally, each word in the input vocabulary would map on to a short phrase. Longer phrases could be built up using combinations of words, meaning each allow sequence of words would produce a unique output sequence (or command).

3.2 STARDUST

The second and third dataset are based on data collected in the STARDUST project [3]. The second dataset is an isolated word recognition task using the same `(sil $word sil)` grammar as the VIVOCA data. It consists of three speakers and is constructed from the available training and adaptation data. The vocabulary size, number of utterances and intelligibility assessment are shown in Table 1.

The third dataset entails command & control sentences. Since the employed databases contain only few, if any sentence recordings we artificially constructed sentences by concatenating the waveforms of isolated words following a speaker-specific grammar. These grammars, shown in Fig. 1, were constructed to closely resemble those used in the STARDUST project, albeit somewhat simplified to account for shortages of some (isolated) words. While not a replacement for the full acoustic variation in real spoken sentences (albeit dysarthric speech may exhibit more pauses between words than regular speech), the data does suffice to evaluate the effectiveness of ALADIN approach of learning without segmentation/word order information.

A Voice Activity Detection (VAD) algorithm [18] was used to remove some of the silence in the isolated word waveforms prior to concatenation, although substantial pre-,inter-,and post-word silence remains. Every isolated word from the second database was (at most) only used once in the construction of the third database. The sentences were randomly generated while maintaining an as even distribution of words and grammar rules as possible. With respect to the isolated words STARDUST database (c.f. Table 1), the vocabulary of speaker 1 changed from 19 to 17 words, and the utterance counts for speaker 1-3 are now 260,204 and 490, respectively.

4 Experimental setup

4.1 ASR frontend

The conventional ASR front-end, referred to as *ASR* in the experimental results, employs left-to-right HMMs with 9 states per word, which yielded slightly better results than the 11 states employed in [5]. Lower state counts were explored as well (not shown), but those lead to only small improvements with few training samples, at the cost of a decrease with more data. The acoustic vectors were 12 Mel-frequency cepstral coefficients (MFCCs) derived from a 26-channel filter-bank with a 25 ms analysis window and 15 ms frame-rate. Energy normalization and cepstral mean normalization was applied to the input features. The models were trained using the HMM toolkit [19] with the Baum-Welch algorithm.

$device1 = tv | disc | radio;
$device2 = film;
$device3 = video;
$state = on | standby;
$control1 = sound | channel;
$control2 = play | stop;
$dir = up | down;
$nums = one | two | three | four | five;
$cancel = no;

$cstate = sil **$device1** sil **$state** sil;
$ccntrl1 = sil **$control1** sil **$dir** sil;
$ccntrl2 = sil **$device2** sil **$control2** sil;
$ccntrl3 = sil **$device3** sil **$nums** sil;
$ccancel = sil **$cancel** sil;

(**$cstate** | **$ccntrl1** | **$ccntrl2** | **$ccntrl3** | **$ccancel**)

(a) STARDUST speaker 1

$device = tv | radio | lamp;
$state = on | standby;
$control = volume | channel;
$dir = up | down;
$cancel = no;

$cstate = sil **$device** sil **$state** sil;
$ccntrl1 = sil **$control** sil **$dir** sil;
$ccancel = sil **$cancel** sil;

(**$cstate** | **$ccntrl1** | **$ccancel**)

(b) STARDUST speaker 2

$device = tv | disc;
$state = on | standby;
$control1 = volume | channel;
$control2 = play | stop | forward | back;
$dir = up | down;
$cancel = bugger;

$cstate = sil **$device** sil **$state** sil;
$ccntrl1 = sil **$control1** sil **$dir** sil;
$ccntrl2 = sil **$control2** sil;
$ccancel = sil **$cancel** sil;

(**$cstate** | **$ccntrl1** | **$ccntrl2** | **$ccancel**)

(c) STARDUST speaker 3

Figure 1: Grammars and vocabulary for each of the three speakers in the STARDUST sentence dataset.

Frame	*Slot*	*Value*
on_off	<action>	on,off
	<device>	1-3
control	<action>	up,down
	<function>	vol,chan
film	<action>	play,stop
video	<action>	1-5
cancel	-	-

(a) STARDUST speaker 1

Frame	*Slot*	*Value*
on_off	<action>	on,off
	<device>	1-3
control	<action>	up,down
	<function>	vol,chan
cancel	-	-

(b) STARDUST speaker 2

Frame	*Slot*	*Value*
on_off	<action>	on,off
	<device>	tv,disc
control	<action>	up,down
	<function>	vol,chan
disc	<action>	1-4
cancel	-	-

(c) STARDUST speaker 3

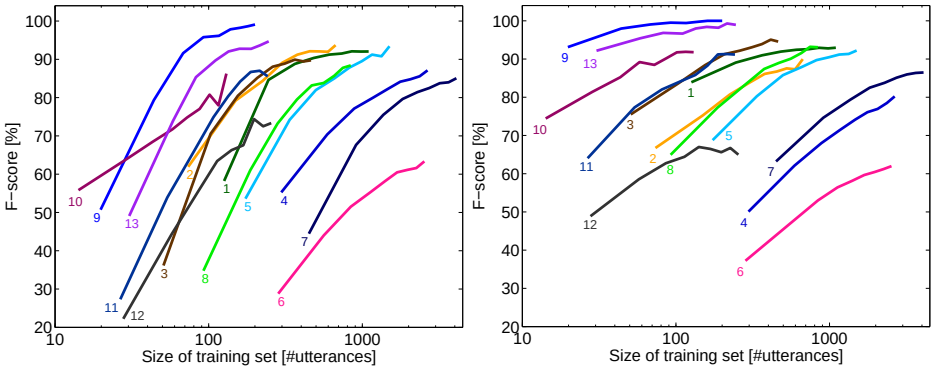
Figure 2: Semantic frame descriptions for each of the three speakers in the STARDUST sentence data. Note that the slot-values do not directly correspond to the vocabulary in the grammars in Fig. 1, as they only represent human-readable tags of semantic concepts.

4.2 ALADIN

The ALADIN system employs a VAD [18] to remove silence prior to feature extraction, and used per-utterance mean & variance normalisation on the extracted MFCC features. The mid-level acoustic representation, unique to each speaker, consists of 100 full-covariance Gaussians, trained on all speech material available for that speaker. For the isolated word experiments, the semantic frame descriptions entail a single (empty) frame per word. The semantic frame descriptions for the sentence data were modelled after the grammars in Fig. 1 and are shown in Fig. 2. Other parameter settings were taken the same as in [11]. Note that on isolated word data, the NMF-based learning boils down to a (Kullback-Leibler divergence weighted) averaging of the co-occurrence acoustic representations for each word.

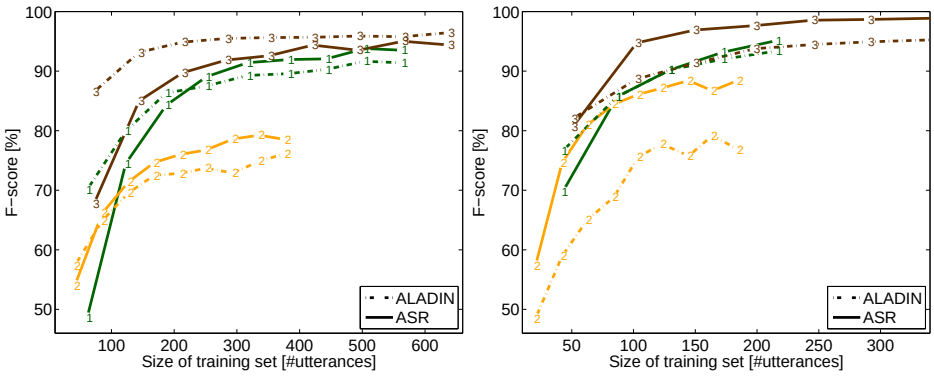
4.3 Evaluation procedure

We use the cross-validation technique described in [11]. In short, we divide the data in multiple blocks, with the constraints that each slot-value should



(a) ASR (b) ALADIN

Figure 3: VIVOCA isolated word recognition results per speaker as a function of the averaged number of utterances in the training set. The left panel displays the results obtained with the ASR system, a conventional GMM-HMM recognizer, whereas the right panel displays the results obtained with the NMF-based ALADIN framework. The graphs are displayed with a logarithmic horizontal axis to account for the large differences in the amount of training material. Numbers indicate the speaker index.



(a) Isolated words (b) Sentences

Figure 4: STARDUST results per speaker as a function of the averaged number of utterances in the training set. The left panel displays the results obtained with isolated word recognition, whereas the right panel displays the results obtained with command & control sentences. Numbers indicate the speaker index.

occur in each block, and that the distribution of slot-values over blocks is as equal as possible. We evaluate with an increasing number of blocks used as training data, with the remaining blocks used as test data. The number of blocks is dependent on the amount of speech material and ranges from 10 to 6. To improve the statistical significance, we repeat the procedure with five different assignments of blocks to train and test data (folds). Evaluation is done using an F-score measure at the slot-value level, aggregated over all five folds. For more details we refer the reader to [11]. Note that for the isolated words datasets, the use of a single frame per word means the F-score is equal

to the word classification accuracy.

5 Results

The results of the evaluation on the isolated words VIVOCA database are shown in Fig. 3. For both the ASR and the ALADIN system, we observe large performance differences between speakers at the end of the learning curves. The speakers with the best intelligibility assessment (2, 9 and 13) are among the best performing, but several other speakers (such as 1 and 5) perform comparably. At the same time, some speakers, such as 6 and 12, do not exceed F-scores of 70-75% even with substantial amounts of training data (hundreds of examples per word).

When comparing the ASR and the ALADIN system, we can observe that the ALADIN system achieves much higher F-scores at the beginning of the learning curves, ranging from 5 to 40% absolute. This remains true even for speakers for which the beginning of the learning curve represents dozens of examples per word — for speakers with much speech material the cross-validation procedure resulted in relatively initial training blocks. At the end of the learning curve, the systems perform comparably for most speakers, and ASR and ALADIN outperforming each other on some. For the isolated words STARDUST dataset in Fig. 4, we observe the same trends.

On the STARDUST artificial sentence data displayed in Fig. 4b the situation is reversed: The ASR approach benefits greatly from the constraints imposed by the grammar, with ASR now performing better than ALADIN even at the beginning of the learning curve for speakers 2 and 3, while performing comparably on speaker 1 at both ends.

6 Discussion and conclusions

Direct comparison with the isolated word dataset is not possible, due to differences in the training size per cross-validation block, the vocabulary size (for speaker 1) and the recognition metric (for the sentence data the F-score is not equal to the word recognition accuracy), That said, it seems reasonable to conclude that the ASR results improve from the additional constraints imposed by the grammar, while the ALADIN results decrease due to the difficulty of learning patterns from utterance-based representations for words that are never seen in isolation.

For isolated words — the usage scenario of the existing VIVOCA system described in VIVOCA — the ALADIN system may already be a viable approach to reduce the amount of training data needed. For sentence data, more evaluation is needed, although it is impressive that the ALADIN approach performs comparably to ASR for STARDUST speaker 1 even though that speaker has the most complex grammar. Future work will focus on comparisons on sentence data from more speakers, real sentences, and with less constrained grammars and vocabulary.

References

[1] D. Beukelman and P. Mirenda, "Augmentative and alternative communication," 2005.

[2] J. Noyes and C. Frankish, "Speech recognition technology for individuals with disabilities," *Augmentative and Alternative Communication*, vol. 8, no. 4, pp. 297–303, 1992.

[3] M. S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. O'Neill, and R. Palmer, "A speech-controlled environmental control system for people with severe dysarthria," *Medical Engineering & Physics*, vol. 5, no. 29, pp. 586 – 593, 2007.

[4] J. F. Gemmeke, B.Ons, N.Tessema, H. Van hamme, J. van de Loo, W. D. G. De Pauw, J.Huyghe, J. Derboven, L.Vugen, B. van Den Broeck, P. Karsmakers, and B. Vanrumste, "Self-taught assistive vocal interfaces : An overview of the ALADIN project," in *Proc. INTERSPEECH*, 2013, pp. 1–5.

[5] M. Hawley, S. Cunningham, P. Green, P. Enderby, R. Palmer, S. Sehgal, and P. O'Neill, "A voice-input voice-output communication aid for people with severe speech impairment," *Neural Systems and Rehabilitation Engineering, IEEE Transactions on*, vol. 21, no. 1, pp. 23–31, Jan 2013.

[6] H. Christensen, S. Cunningham, C. Fox, P. Green, and T. Hain, "A comparative study of adaptive, automatic recognition of disordered speech," in *Proc Interspeech 2012*, Portland, Oregon, US, Sep 2012.

[7] K. T. Mengistu and F. Rudzicz, "Comparing humans and automatic speech recognition systems in recognizing dysarthric speech," in *Proceedings of the Canadian Conference on Artificial Intelligence*, 2011.

[8] H. V. Sharma and M. Hasegawa-Johnson, "State transition interpolation and map adaptation for HMM-based dysarthric speech recognition," in *HLT/NAACL Workshop on Speech and Language Processing for Assistive Technology (SLPAT)*, 2010, pp. 72–79.

[9] F. Rudzicz, "Acoustic transformations to improve the intelligibility of dysarthric speech," in *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT2011)*, 2011.

[10] J. F. Gemmeke, J. van de Loo, G. De Pauw, J. Driesen, H. Van hamme, and W. Daelemans, "A self-learning assistive vocal interface based on vocabulary learning and grammar induction," in *Proc. INTERSPEECH*, 2012, pp. 1–4.

[11] B. Ons, N. Tessema, J. van de Loo, J. F. Gemmeke, G. De Pauw, W. Daelemans, and H. Van hamme, "A self learning vocal interface for speech-impaired users," in *Proc. Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2013, pp. 73–81.

[12] Y. Wang and A. Acero, "Rapid development of spoken language understanding grammars," *Speech Communication*, vol. 48, no. 3-4, pp. 390–416, 2006.

[13] B. Ons, J. F. Gemmeke, and H. Van hamme, "Label noise robustness and learning speed in a self-learning vocal user interface," in *Proc. of the International Workshop on Spoken Dialog Systems (IWSDS)*, Ermenonville, France, 2012.

[14] D. Lee and H. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[15] J. Eggert and E. Korner, "Sparse coding and NMF," in *Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on*, vol. 4, 2004, pp. 2529–2533 vol.4.

[16] Y.-X. Wang and Y.-J. Zhang, "Nonnegative Matrix Factorization: A comprehensive review," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 25, no. 6, pp. 1336–1353, 2013.

[17] H. Lee, J. Yoo, and S. Choi, "Semi-supervised nonnegative matrix factorization," *Signal Processing Letters, IEEE*, vol. 17, no. 1, pp. 4–7, 2010.

[18] J. Ramírez, J. M. Górriz, J. C. Segura, C. G. Puntonet, and A. J. Rubio, "Speech/non-speech discrimination based on contextual information integrated bispectrum LRT," *Signal Processing Letters, IEEE*, vol. 13, no. 8, pp. 497–500, 2006.

[19] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The htk book," *Cambridge University Engineering Department*, vol. 3, 2002.